SUPPORTING SECONDARY USE OF DATA: THE ORD EXPERIENCE USING THE ENVIRONMENTAL INFORMATION MANAGEMENT SYSTEM (EIMS)

Linda Kirkland, Ph.D.
Environmental Scientist
EPA/NCERQA
401 M Street, SW
Washington, DC. 20460

Jeffrey B. Frithsen, Ph.D.
Info. Management Specialist
EPA/NCEA
401 M Street, SW
Washington, DC 20460

Robert F. Shepanek, Ph.D.
Info. Management Specialist
EPA/NCEA
401 M Street, SW
Washington, DC 20460

SUMMARY

The nature of environmental assessments has changed placing increasing emphasis on the integration of data collected by multiple investigators and representing multiple disciplines. In response, information management programs are developing approaches to promote the sharing of data and descriptive information about data across distributed networks. A particular technical and management challenge is ensuring that descriptive information about data (metadata) is captured and is sufficient to enable investigators to evaluate the application of the data for a use other than that for which the data were collected. The Environmental Information Management System (EIMS) is being used to capture, store, manage, and provide data and metadata to users within the Office of Research and Development, the Office of Water, Region 10, and the public. This system is described and an example of metadata for a data set provided as part of steps leading to developing scientific metadata standards for ORD. These standards would require content such as descriptive information on data quality produced by the EPA's Quality Assurance Program [e.g., proficiency testing (QC), data validation quality indicators (DQI), and data quality assessment (DQA) results].

INTRODUCTION

Increasingly, the completion of environmental assessments requires integration of data collected by multiple investigators and representing multiple disciplines and multiple spatial and temporal scales (Brown 1994). One example of the complexity of these assessments is reflected in a recently proposed national environmental monitoring framework (NSTC 1997). This framework suggests that better assessment of environmental condition and environmental trends can be completed through the integration of remote sensing, regional monitoring, and site intensive studies.

The increased emphasis on multiple investigator research requiring shared access to data presents several information management challenges. Those challenges include the need to:

- Facilitate identifying and finding environmental data and information through the development of a directory of inventory

- Provide better access to descriptive information about environmental resources to facilitate assessments of secondary use
- Promote the sharing of environmental resources within teams of investigators distributed geographically across organization boundaries
- Expand the distribution of data and information to the public.

The Environmental Information Management System (EIMS) serves as a model of a system that begins to meet these challenges. This system evolved from the information management system developed for the Office of Research and Development's Environmental Monitoring and Assessment Program (EMAP) (Shepanek 1994). The continued development and growth of EIMS has been led by the ORD's National Center for Environmental Assessment (NCEA). NCEA is responsible for the development of ecological and human health risk assessment methodologies and guidelines, and provides support to EPA Program Offices and Regions conducting risk assessments. Thus, NCEA is most often a secondary user of data and is positioned to provide guidance on the development of scientific information management systems needed to complete environmental assessments (USEPA 1997).

EIMS was designed to capture, store, and manage data and descriptive information about data. This descriptive information is frequently referred to as metadata, literally data about data. At one level, metadata help users to identify and locate data much like a library catalog helps to identify and locate documents. At another level, metadata are used to provide a detailed understanding of the data enabling a user to evaluate if data can reasonably be used in ways not originally intended by the data originator. EIMS was designed to include metadata needed to fulfill both functions.

EIMS is evolving to include a robust set of metadata; however, guidelines and procedures will need to be developed to define a minimum set of metadata that may be required to ensure that adequate documentation is available to assess secondary use of the data. The purpose of this paper is to show how the metadata for a data set are organized within EIMS, provide a metadata example for a data set, and present a summary of minimum requirements for scientific metadata.

SYSTEM DESCRIPTION

The information management capabilities required for complex environmental assessments are reflected in the components of EIMS (Figure 1). The main components are the metadata, represented by the directory, catalog, and dictionary, and the data. The EIMS directory contains information about "objects" that are relevant to the assessment process. These objects include projects, documents, data sets, databases, and software tools. The types of information that the directory contains about objects include a short narrative abstract, contact information, locational information, and keywords that allow the directory to be searched in a variety of ways. The function of the directory with respect to assessment is analogous to a bibliographic search

for references prior to writing a scientific paper.

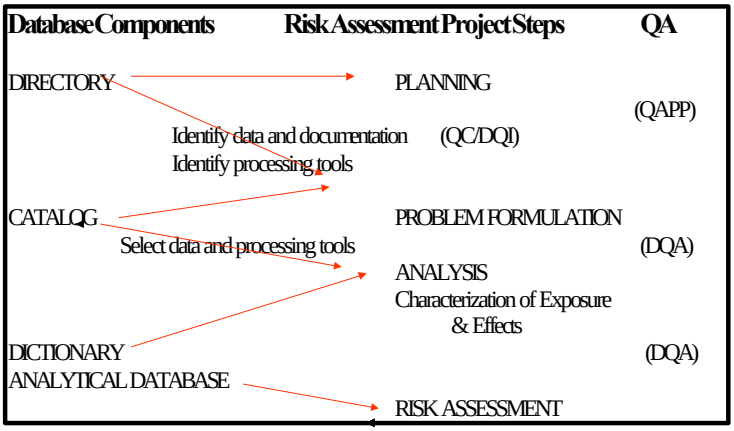

Figure 1. EIMS components.

Once environmental resources are located, an evaluation must be made concerning their potential use. The EIMS catalog is designed to store and organize detailed information about objects that are indexed in the directory. The level of detail is sufficient to allow an assessment scientist to make a determination of whether a particular data set, database, or tool is appropriate for an intended use. Examples of information in the catalog are definitions of individual attributes in a data set, information about methods used to collect and analyze the data, and quality assurance information. The catalog provides a template to guide and capture information about data that are produced during the process of evaluating data for secondary use.

Data that are judged to be useful can be downloaded and used; however, actual use of data requires fundamental information about individual elements within a data set (for example, the name of the variable containing dissolved oxygen measurements). The dictionary contains specific information about each of the attributes in a data set or database. If a data set is included in the database, the dictionary is the means by which metadata in the directory and catalog are linked to the information that is included in the database structure. Examples of information in the dictionary are the format, length, definition, and allowable values for a specific attribute.

The final component of the EIMS design, is the data. Not all data sets need be loaded into the database; however, in those cases where several investigators are working with the same data the content of which is changing with some frequency, or for which the same types of analyses are needed repeatedly, the development of a consistent analytical database is appropriate. The purpose of the EIMS analytical database is to facilitate the process of analyzing multiple data sets together. Initial selection of data sets for a particular assessment project is made by reviewing directory information. Data sets of interest then are evaluated using the catalog as a template. Those that are still deemed valuable in support of the assessment are loaded into the analytical component of the database. This process results in a collection of data organized in a homogeneous way, which greatly facilitates access and manipulation, using a variety of analytical tools. Examples of tools are geographic, visualization, statistical, and spreadsheet systems.

The components of EIMS are linked enabling users to use the directory to find data in the system, or to find a particular measurement method associated with a specific data value in the

database. This linkage of metadata and data is invaluable when data are integrated from multiple sources, each having slightly different methods or measurement protocols.

METADATA COMPLETENESS

The metadata components of EIMS were developed to provide an inventory of environmental resources and to support assessments of secondary use. The components providing the inventory, or library card catalog function were based upon metadata components developed for the NASA directory interchange format (NASA 1991). The components providing a detailed description of the data were based upon various examples, including the detailed documentation compiled for NASA's FIFE Project (Strebel et al. 1990 a,b) and the spatial metadata standards developed by the Federal Geographic Data Committee (FGDC 1997).

The content of organization of EIMS metadata is designed to meet the 20-year rule proposed by the National Research Council's Committee on Geophysical Data: " Will someone 20 years from now, not familiar with the data or how they were obtained, be able to fine data sets of interest and then fully understand and use the data solely with the aid of the documentation archived with the data set?" (NRC 1991).

The 20-year rule guideline for meta-data is also consistent with the recently proposed USEPA-Office of Research and Development's (ORD) policy that its research data under go quality assurance review before release and that records must be retained in sufficient detail so that individuals trained in the appropriate disciplines can reconstruct the research. Quality assurance procedures produce records that describe what the data represents and how well they do it (Figure 2), and the results of those procedures need to be captured as part of the metadata associated with the data set. Ultimately, these scientific meta-data also provides the basis for variability and uncertainty analyses supporting risk assessments (ORD1997).

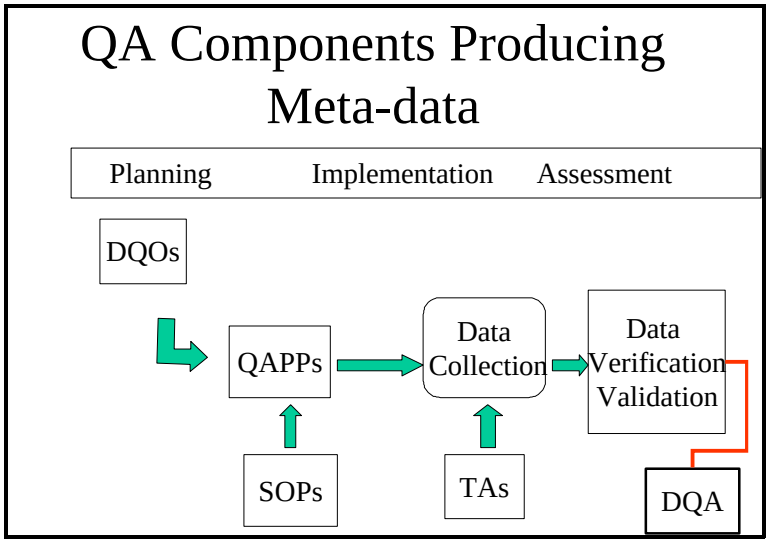

Figure 2: Metadata quality assurance components.

METADATA RECORD CONTENT

A complete metadata record contains directory, catalog, and dictionary components. A representative list of the individual data fields included in a metadata record are given in Table 1. The list is not complete and is meant only to give an overview of the types of information needed to compile a complete metadata record satisfying the 20-year rule. The information is organized into sections reflecting the typical organization of a scientific paper. For example, there are sections for title, authors, abstract, introduction and objectives, and methods. In place of a description of results as would be present in a scientific paper, the metadata record contains a description of data.

CONCLUSIONS

There are several lessons learned from the development and implementation of EIMS pertaining to the compilation of scientific metadata sufficient to support secondary use of data. These lessons are summarized below.

1. Undocumented data are unuseable data. Data documentation is needed to assure that data collected and used by the agency can be used for future studies.

2. The compilation of summary level metadata reflective of the contents of the EIMS directory is relatively easy to do, does not require large resources, and will assist users with identifying and finding data; however, summary level metadata will not assist with evaluations of data for secondary use.

3.. The compilation of a complete metadata record sufficient to meet the 20-year rule requires a significant amount of time.

4. Metadata compilation needs to be completed by the scientists who have collected the data. Data documentation cannot be completed successfully by information management scientists.

Table 1. Representative list of data fields included in a complete metadata record for a data set.

I. Data Set Identification Fields
- Title of data set
- Data set version
- Date of metadata entry
- Metadata author
- Metadata record review status

II. Investigator Information Fields
- Principal investigator
- Sample collection investigator
- Sample processing investigator
- Data analysis investigator

III. Abstract Fields
- Abstract
- Descriptive keywords for subject

IV. Objectives Fields
- Program objective
- Data set objective

V. Data Acquisition Method Fields
- Sample collection method summary
- Sampling platform
- Sampling equipment
- Sampling method calibration
- Sample collection quality control
- Sample collection method reference
- Sample collection method deviations

VI. Sample Processing Method Fields
- Sample processing method summary
- Sample processing method calibration
- Sample processing quality control
- Sample processing method reference
- Sample processing method deviations

VII. Methods for Derived Data
- Name of modified or derived data element
- Data derivation method description
- Data derivation examples
- Data derivation computer code
- Computer code language
- Computer cod file name

VII. Data Element Description Fields
- Data element name
- Data element description
- Data element type and format
- Allowable minimum values in data set
- Allowable maximum values in data set
- Description of criteria for allowable data range

VIII. Quality Control/Assurance Information Fields
- Measurement quality objectives
- Quality assurance/control method descriptions
- Reported measurement quality
- Identified sources of error
- Confidence level/accuracy judgement
- Quality assurance reference data
- Comments on data use and constraints

IX. Geographic and Spatial Information Fields
- Minimum and maximum latitude
- Minimum and maximum longitude
- Minimum and maximum depth or altitude
- Horizontal coordinate system
- Resolution of horizontal coordinates
- Horizontal coordinate accuracy
- Vertical coordinate system
- Resolution of vertical coordinates
- Vertical coordinate accuracy

X. Data Access Information Fields
- Data access procedures
- Data access restrictions
- Data access contract person
- Data set formats
- On-line access information

XI. Reference Fields
- Suggested reference for metadata record
- Requested acknowledgment
- Related references

REFERENCES

Brown, J.H. 1994. Grand challenges in scaling up environmental research. In, W.K. Michener, J.W. Brunt, and S.G. Stafford (eds.) Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor & Francis, Bristol, PA. pp. 21-26.

Federal Geographic Data Committee. 1997. Content standard for digital geospatial metadata (revised April, 1997) Federal Geographic Data Committee. Washington, DC.

NASA 1991. Directory Interchange Format Manual. Version 4.0, December 1991. NASA, National Space Science Data Center, Greenbelt, MD.

NRC 1991. Solving the global change puzzle: A U.S. Strategy for managing data and information. A report by the Committee on Geophysical Data, Commission on Geosciences, Environment, and Resources, National Research Council. National Academy Press, Washington, DC.

NSTC 1997. Integrating the Nation's environmental monitoring and research networks and programs: A proposed framework. March 1997. National Science and Technology Council, Committee on Environment and Natural Resources, Environmental Monitoring Team. Washington, DC.

Office of Research and Development, Risk Assessment Forum. 1997. Guiding Principles for Monte Carlo Analysis. EPA/630/R-97/001.

Shepanek, R. 1994. EMAP Information Management Strategic Plan: 1993-1997. EPA/620/R-94/017. Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Assessment Program.

Strebel, D.E., et al. 1990. The FIFE information system: Support of interdisciplinary science. In, Symposium on the First ISLSCP Field Experiment, American Meteorological Society, Boston, pp. 140-147.

Strebel, D.E., J.A. Newcomer, J.P. Ormsby, F.G. Hall, and P.J. Sellers. 1990. The FIFE Information System. IEEE Transactions on Geoscience and Remote Sensing 28: 703-710.